Gaining Weights... and Feeling Good about It!

Wit, Ernst; Purutcuoglu, Vilda; O'Donovan, Lucy; Zhu, Ximin

Published in:
EPRINTS-BOOK-TITLE

Publication date:
2007

CHAPTER 4

Gaining Weights ... and Feeling Good about It!

Ernst Wit,* Vilda Purutcuoglu, Lucy O'Donovan and Ximin Zhu

Abstract

Two problems that dog current microarrays analyses are (i) the relatively arbitrary nature of data preprocessing and (ii) the inability to incorporate spot quality information in inference except by all-or-nothing spot filtering. In this chapter we propose an approach based on using weights to overcome these two problems. The first approach uses weighted p-values to make inference robust to normalization and the second approach uses weighted spot intensity values to improve inference without any filtering.

A Light Introduction

As with many other types of high-throughput technologies, microarray data require essential preprocessing steps in order to present it in a format that can be used for making inference. From the moment the actual experimental procedure have been completed after the hybridization a combination of several crucial steps have to be undertaken in order to get data. First of all the slides are scanned, which turns the number of attached mRNA molecules into collection of pixel values within an image. Then an image analysis package separates the background from the foreground signal (**gridding** and **segmentation**) and combines the pixel values into range of summaries (quantification). Those summaries typically consist of quantities like the mean, the median and the variance of the spot as well as the background pixel values. However, of those summaries, typically only a single value, namely the spot mean or median is used in inference. In section 3 we shall deal with ways we can use more of the available outputs in inference.

Those spot values are then frequently **normalized** across probe sets, array, channels or the whole experiment often changing the scale of the data via a number of possible algorithms, usually combined in some computer package (e.g., MAS 5.0, RMA, smida). This preprocessing of the physical, hybridized slides $S = (S_1, \ldots, S_s)$ into a data matrix of gene expression values $Y = (Y_1, \ldots, Y_s)$ can be represented via the action of the operator f

$$Y = f(S, \nu),$$

where, crucially, the ν stands for all the parameters and normalization settings used in the preprocessing. In turn, these parameters are intended to capture the physical process of turning mRNA counts into an image into gene expression values. The precise value for ν is typically unknown and depends highly on the skill of the technicians and software involved in the preprocessing steps for what seem **reasonable** choices. It is rare, although not impossible,

*Corresponding Author: Ernst Wit—Medical Statistics Unit, Department of Mathematics and Statistics, Lancaster University, Lancaster, LA1 1AE U.K. Email: e.wit@lancaster.ac.uk

Microarray Technology and Cancer Gene Profiling, edited by Simone Mocellin.

that a value for ν can be estimated from the data. Therefore, there is typically some level of arbitrariness in the choice of ν. Slightly different values of ν, e.g., ν*, will lead to different values of Y,

$$Y^* = f(S, \nu^*).$$

Which value should we actually use in our analysis? Y or Y^*, or perhaps some completely different Y^{**}. Most practical bioinformaticians would probably feel that they could live with this situation as long as they feel that they have made a "reasonable" choice for ν. In that case, they would calculate, for example, their t-statistic $t_g(Y_g)$ for a particular gene g on the basis of the available data Y_g and calculate the two-sided p-value as

$$\text{p} - \text{value}(\nu) = 2\text{p}_{H_{0,\nu}}\left(\text{T} > \left|t_g(Y_g)\right|\right),$$

where H_0 is the null-hypothesis of no differential expression and, importantly, ν the actually selected preprocessing parameters.

However, it is possible that different reasonable ν-values will give rise to different answers, such as different significantly expressed genes—had one only tried. In section 2 we deal with the simple question to what extent we can accommodate the actual level of arbitrariness in the preprocessing of the data within our inference.

P-Value Weighting

If we have control over at least some of the nuisance parameters ν, it is in principle possible to vary them to study their effects on inference. Consider for example that we could vary the gain settings on the scanner, the morphological properties of the image analysis programme or the parameters of the normalization procedures, which in total represents m different parameters, $\nu = (\nu_1, \ldots, \nu_m)$.

"Reasonable" values for ν can be expressed as a hypothetical distribution on the parameters, p_ν. This distribution expresses all the uncertainty about the preprocessing process—much in a way a Bayesian prior distribution would do. This uncertainty about ν propagates into uncertainty about the data Y, which in turn modifies, for example, the p-value for gene g,

$$p - value(g) = P_{H_0}\left(T > \left|t_g(Y_g)\right|\right) = \int P_{H_{0,\nu}}\left(T > \left|t_g(Y_g)\right|\right) p_\nu(\nu)\partial\nu. \quad (1)$$

This is the real p-value, i.e., the p-value that takes into account all the uncertainty about the data. In other words, the real p-value is a weighted average of all the naïve p-values at a particular normalization setting.

What does this mean in practice? As the normalization procedures can be extremely complex, it is unlikely that the integral in (1) can be solved explicitly. As a result, numeric integration via a discrete sum is the only way to make progress. In particular, if $N = \{\nu^1, \ldots, \nu^k\}$ is the set of k normalization settings with weights $w_1, \ldots, w_k$, giving rise to k alternative data sets $\{Y^1, \ldots, Y^k\}$, then an approximate p-value can be calculated as

$$p - value(g) = \sum_{\nu \in N} w_\nu P_{0,\nu}\left(T > \left|t_g(Y_g)\right|\right) / \sum_{\nu \in N} w_\nu \quad (2)$$

Effectively, the distribution $w_{\nu 0}/\Sigma_\nu w_\nu$ is the discretised version the normalization parameter distribution. The true p-value is therefore a weighted sum of the p-values corresponding to the individual normalizations.

For good measure, we should add that in the presence of control spots on the microarray, the values of at least some of the preprocess parameters could be estimated directly from the data. This means that the subjective distribution $p\nu$ can be replaced by an objective distribution, which now represents the uncertainty in the estimates of ν. In case many of such control spots are present on the microarray and ν can be estimated quite precisely, then inference can be done on the single normalized dataset where $\nu = \hat{\nu}$.

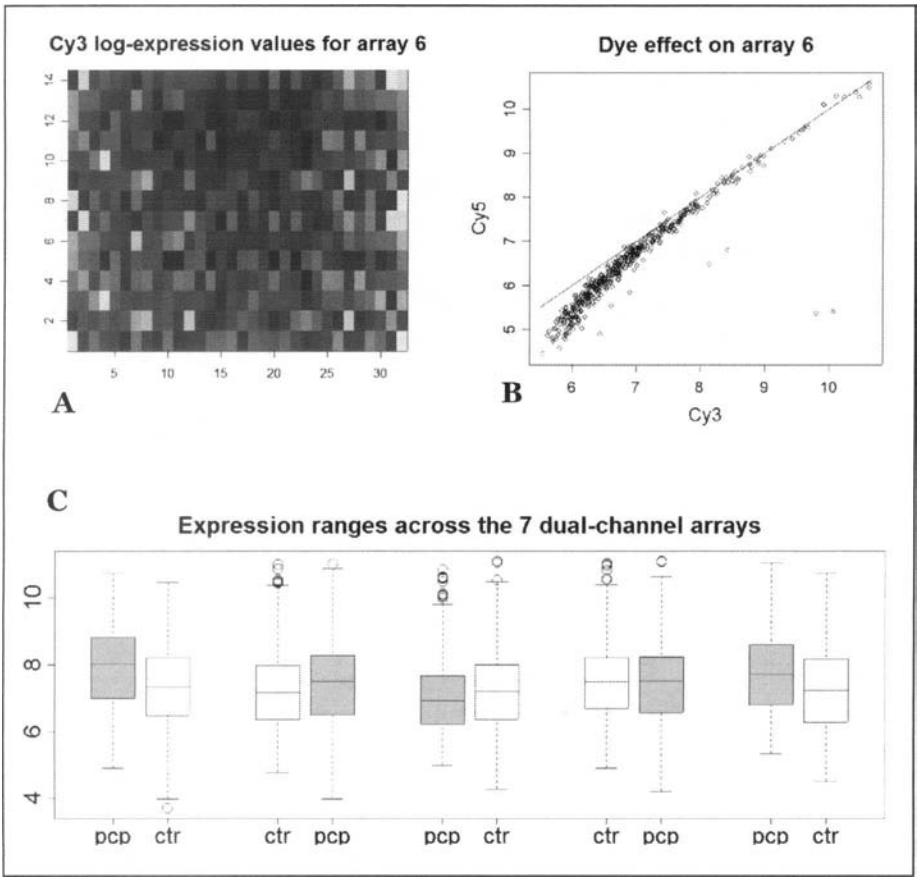

Figure 1. Some of the typical artifacts present in the mouse-PCP experiment: A) a spatial effect; B) a uneven dye effect; and C) a cross-comparison issue over the 7 arrays.

Application

Dr. Lucy O'Donovan (University of Glasgow) performed a microarray experiment, in which one of the aims was to find those genes that are differentially expressed in a mouse schizophrenia model as compared to in wild-type mice. The schizophrenia model was induced by treating the mice with a drug, PCP. Dr. O'Donovan hybridized the RNA from seven PCP and seven wild-type mice in a pairwise fashion to seven dual-channel microarrays. Each of the arrays contained 224 genes, spotted in duplicate. Although quality control measures suggested seven good hybridizations, with some transformation of the data several artifacts of the data were easy to spot by eye. In particular, there was some uneven hybridization across the arrays with darker areas in the top centre part of the array (Fig. 1A), some uneven dye effects (Fig. 1B) and also uneven gains between the arrays as evidenced by (Fig. 1C). In order to deal with these nuisance effects, we applied a series of preprocessing steps contained in the R-package smida. The mere default settings of the normalization procedures resulted in visually "improved" data (Figs. 2A-C).

The spatial normalization was done by fitting a first degree *loess* curve to both the mean and the standard deviation of the log-transformed data with span parameters equal to 0.5

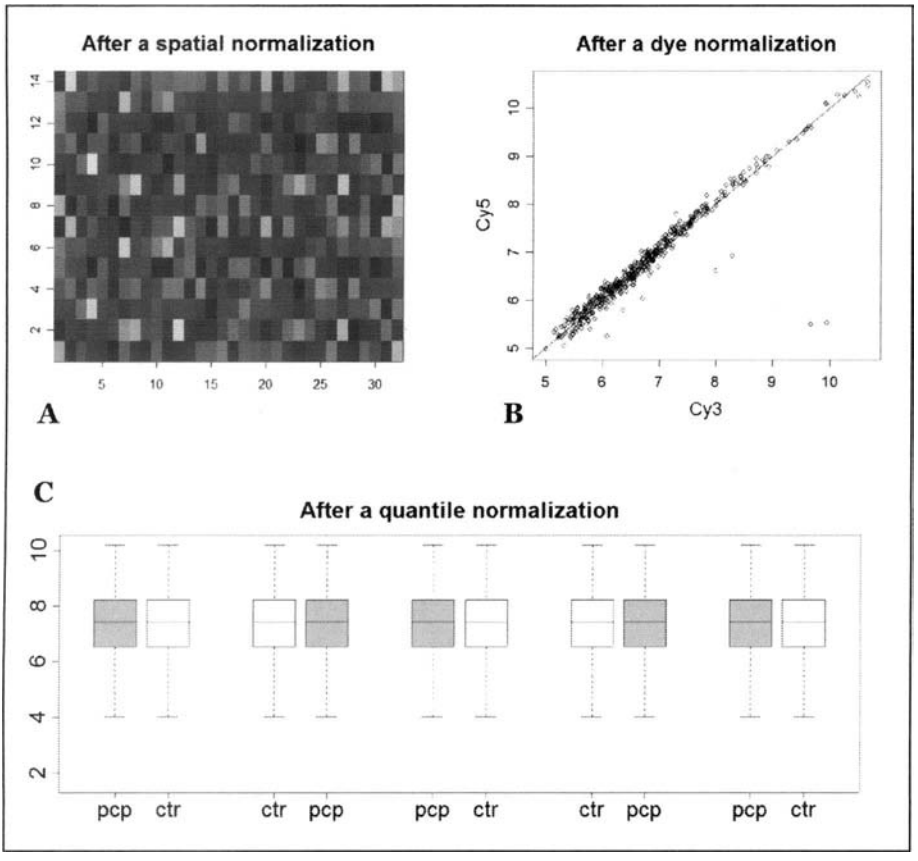

Figure 2. Preprocessing of the data helps to overcome: A) a spatial effect by brightening the top-centre part of the array; B) a uneven dye effect by boosting the Cy5 values in the lower expression ranges; and C) a cross-comparison issue by rescaling all the arrays to the same average distribution.

and 0.75 respectively. Changing the parameter values for the location normalization to 0.2 has no obvious visual effect on the normalization. We also consider a scale span parameter of 0.3 instead of 0.75.

By default the smida package does not subtract background, however settings are available to do either a probabilistic or deterministic background subtraction (details in ref. 1, sect. 4.3.3). The default setting for the *loess* dye normalization is a span of 0.2. By changing this span to 0.5 the adjustment becomes slightly less variable across the intensity range, although the effect is almost invisible. Similarly, for the quantile normalization one needs to specify a set of invariant genes. Complete quantile normalization implicitly assumes that all genes are invariant. We also considered invariant set sizes of 30 and 100 genes out of all 224 genes. Taken altogether, we considered

$$2 \text{ spat loc} \times 2 \text{ spat scale} \times 3 \text{ bkg} \times 2 \text{ dye} \times 3 \text{ quantile} = 72 \text{ normalizations}$$

Each of these normalizations resulted in an alternative normalized dataset. In each of these datasets, we could proceed to test for differential expression across each of the 224 genes. Standard normal quantile plots suggest that the normal assumption is not inappropriate and that therefore a t-test can be used to find differences between PCP and control mice.

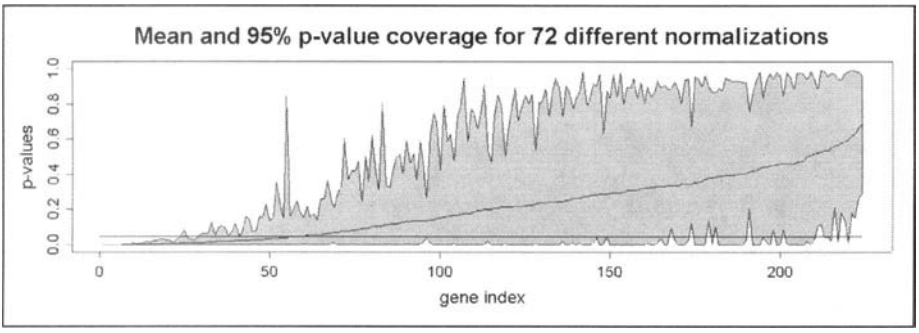

Figure 3. For each of the 72 normalizations and for each of the 224 genes we calculate all of the p-values for testing for differential expression between PCR and control mice. For each of the genes, we indicate the range between the 3rd smallest and 70th largest p-value (approx. 95% coverage), as well as the average p-value.

Figure 3 shows the range of p-values across 72 normalizations for the 224 genes. The genes are arbitrarily ordered by an increasing average p-value, indicated by the solid line in Figure 3. First of all, it is striking to see the impact of the preprocessing on the actual inference from the data.

As we a priori do not have any information to suggest which of the normalizations is better, we regard each of the 72 parameter settings as equally plausible, i.e., $p_V(v) = 1/72$. Consequently, the p-value that is robust to preprocessing is simply given as the average p-value across all 72 normalizations. From Figure 3, we see that 23 genes have a p-value less than 0.01, 59 genes have a p-value less than 0.05. If we use the Benjamini and Hochberg[2] procedure, we find 88 genes that such that the false discovery rate is less than 5%. Clearly, if there had been any control spots on the array for which we could get some idea of the relative plausibility of the normalization parameters, then the relative weights and therefore the resulting p-value would change. Nevertheless, inference based on the average p-value based on several normalizations has the distinct advantage of making inference less susceptible to some arbitrary settings.

Within-Spot Pixel Variance Weighting

That not every gene expression measurement is as good as another is well known in the microarray community. Soon after the introduction of microarrays, imaging programmes introduced the idea of spot filtering: a method whereby unreliable spots were flagged in order to remove them from analysis. Even in its crudest form, flagging is an example of 0-1 weighting. However, there is no reason why these all-or-nothing weights cannot be replaced by more realistic, continuous weights. In this section, we derive a simple method for introducing continuous weights.

In this section we assume that the quality of a spot can be indicated by means of a single value, which corresponds to the within-spot pixel standard deviation. Highly variable spots have a large within-spot pixel standard deviation, whereas good quality, homogeneous spots have small within-spot pixel standard deviations.

Let x_i stand for the ith spot intensity, associated with one particular gene. If we have n spots associated with the same gene, then we can write the spot standard deviations as proportional to the within-spot pixel standard deviation,

$$SD(x_i) \propto \sigma_i, \qquad i = 1,\ldots,n$$

where the constant of proportionality depends on the number of pixels in a spot, the spatial correlation between the spot pixels, which we assume constant across different spots.

If the aim of the experiment is to estimate the true mean expression μ for that particular gene as accurate as possible, then the best estimate can be written as a weighted mean of all the observed expression values,

$$\hat{\mu} = \sum_{i=1}^{n} w_i x_i.$$

If none of the expression values displays any particular bias, then in order for the estimate to be unbiased, the weights should add up to one, $\sum_{i=1}^{n} w_i = 1$. In order to minimize the variance of the estimate, $V(\hat{\mu}) \propto \sum_{i=1}^{n} w_i^2 \sigma_i^2$, it is easy to show that the optimal choice of weights should be proportional to the inverse of the pixel variance in each of the spots,

$$w_i = \frac{1/\sigma_i^2}{\sum_{k=1}^{n} 1/\sigma_k^2}. \tag{3}$$

Although weighting with the within-spot pixel standard deviation is a good idea, there are a few issues that remain to be solved: (i) as the within-spot pixel standard deviations are not known, they have to be estimated from the data; (ii) if the analysis is done on the logarithmic scale, then an appropriate estimate of the within-spot log pixel variance needs to be produced.

Dealing with an Unknown within-Spot Pixel Variance

In order to use the weighting formula in equation (3), we need to replace the unknown within-spot pixel variance with some estimate thereof. The typical output of an imaging and segmentation programme will provide an estimate of the within-spot variance, $\hat{\tau}_i^2$. As we can expect a large amount of spatial correlation between neighbouring pixels, it is likely that this quantity is a severe underestimate of the within-spot pixel variance. Nevertheless, this affects each spot more or less equally and therefore the approximately multiplicative constant should disappear from equation (3).

The most straightforward thing to do would be to replace the within-spot pixel variance σ_i^2 with the estimated within-spot variance $\hat{\tau}_i^2$ in equation (3). However, there may be good reasons to consider a slightly more general approach, namely:

$$\sigma_i^2 \longleftarrow \hat{\tau}_i^2 + a. \tag{4}$$

Taking $a = 0$ is equivalent to using the within-spot variance, whereas if $a \to \infty$ then this would correspond to an unweighted or simple average approach. The reason why adding a constant may have some benefit is that the estimated within-spot variance is itself subject to variation and this constant would robustify the weighting somewhat.

Bakewell and Wit[3] propose a biologically motivated choice for the constant a by replacing it with the estimate of the inherent biological variation for any sample—or in statistical terms, the **subject effect**. Other, more ad hoc choices for a can include some quantile of the observed within-spot variances $\hat{\tau}_i^2$; the higher the quantile, the more conservative and robust the estimate of the weighted mean.

Analysis of the Data on a Log Transformed Scale

Some people prefer to do the analysis microarray data on the logarithmic scale.[4,5] The reason is that the spot intensities are per definition positive, so that it is quite likely that any dominant effects will be multiplicative, rather than additive. Log transforming multiplicative effects will transform the data to an additive scale, which makes analysis more straightforward.

However, most summary information from imaging files is—still—provided on the original pixel scale. Although it would be very useful to have information such as the mean and variance of the log pixel spot values, such information is typically not available. Something can be done, however, in the absence of such information. As is already common, instead of the mean of the log of the spot pixel values, one can use the log of the mean of those values. In

order to get some approximation of the variance of the log pixel values, we can consider the following first order Taylor approximation around the spot mean,

$$V(\log(X)) \approx V\left(EX + (X - EX)\frac{1}{EX}\right) = \frac{V(X)}{E^2(X)}.$$

This equality means that we can approximate the variance of the log pixel values by the original spot pixel variance divided by the square of the original spot pixel mean. An easy way to proceed, therefore, would be as follows:

1. Take as expression values $x_1, \ldots, x_n$,
   x_i = *log spot mean*
2. Take as robust expression variances $s_1^2, \ldots, s_n^2$,
   s_i^2 = *spot variance/(spot mean)*^{2} + *a*
3. Take as weights $w_1, \ldots, w_n$,
$$w_i = \frac{1/s_i^2}{\sum_{k=1}^{n} 1/s_k^2}.$$
4. Calculate the average expression as
$$\hat{\mu} = \sum_{i=1}^{n} w_i x_i.$$

This estimate of the true of expression μ can then be used for further inference, for example, in a test for differential expression. Unfortunately, normal theory does not apply to weighted means, especially for few replicates. Alternatives, such as bootstrap and permutation tests are however directly applicable.

A Weighty Discussion

Preprocessing the data can have a substantial impact on further inference. It is a mistake to assume that the preprocessed data is a unique abstraction of the actual dataset, uniquely suitable for inference. In fact, preprocessing is itself a form of data analysis, which carries along with it all the uncertainty of choosing the correct settings for the normalization parameters. We presented **p-value weighting** as a method to overcome some of the trouble of basing one's conclusions on a single preprocessed dataset.

The method has two main drawbacks. First of all, it requires substantial computational efforts to generate a large number of preprocessed datasets on which to perform the same analysis. As all of the computations are completely parallel, there may be some gain in using parallel computing facilities to overcome some of the computational effort. Even if the computational issue can be overcome, the method is only applicable to hypothesis testing. Only p-values can be "averaged out" over the different datasets, whereas estimates, clusterings or predictions cannot. Nevertheless, the idea of applying the same analysis technique, be it clustering, prediction or something else, to different preprocessed datasets can certainly provide valuable insights. If a particular clustering is stable under the vast majority of all the possible normalizations, then this will strengthen our confidence in the conclusions.

The idea of **spot pixel variance weighting** is sensible because it uses more of the information available in the data, which should lead to better answers. In this sense weighting is similar to data filtering methods. However, spot weighting avoids getting missing values, which is a drawback of most filtering methods.

Should spot weighting be used everywhere? In general, the answer is *yes*. There is valuable information in the output of most imaging and segmentation software. Ignoring such information is equivalent with habitually throwing away a few arrays. Even in cases where the spot pixel variances are slightly suspect or based on only a small number of pixels, combining the observed values with a robustifying choice of constant *a* is expected to improve inference.

Nevertheless, there are a few cases in which using spot pixel variance might be inappropriate: (i) if the quality of a spot is seriously affected by things others than measured in the spot variance; (ii) if the physical meaning of the spot variance changes between arrays. An example of the former is the quality of the RNA used for hybridizing different arrays or the presence of smears and stains, which all may have no influence on the spot variance as such, but does affect the quality of an individual measurement. An example of the latter is a large difference in the gain or large differences in the number of pixels used for spots between different arrays.

References

1. Wit E, McClure JD. Statistics for microarrays: Design, analysis and inference. Chichester: J Wiley and Sons, 2004.
2. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J Royal Statistical Society B 1995; 57:289-300.
3. Bakewell D, Wit E. Weighted analysis of microarray gene expression using maximum-likelihood. Bioinformatics 2005; 21(6):723-9.
4. Irizarry RA, Gautier L, Cope LM. An R package for analyses of Affymetrix oligonucleotide arrays. In: Parmigiani G, Garrett ES, Irizarry RA, eds. The Analysis of Gene Expression Data. Statistics for Biology and Health. New York: Springer-Verlag, 2003:102-19.
5. Kerr M, Churchill GA. Experimental design for gene expression microarrays. Biostatistics 2001; (2):183-201.